RESPONSE LATENCY AS A SIGNAL TO QUESTION PROBLEMS IN SURVEY RESEARCH

JOHN N. BASSILI
B. STACEY SCOTT

Abstract This research explores the potential utility of response latency as an index of question problems in survey research. The time respondents took to answer three types of bad questions was compared to the time they took to answer the repaired versions of the questions. Questions containing a superfluous negative and double-barreled questions took longer to answer than nearly identical questions without these problems. Repaired versions of questions soliciting frequency estimates, however, took longer to answer than their problematic counterparts. The results are discussed in the context of a model of question answering, and their implications for survey methodology are explored.

Survey methodologists have expressed considerable interest recently in techniques for screening survey questions with the aim of repairing bad questions before presenting them to large numbers of respondents (see Presser and Blair 1994). A number of approaches for the early identification of question problems have been explored. Observational monitoring focuses on the interaction between the interviewer and respondent and relies on a behavior-coding scheme to identify problems (e.g., Fowler and Cannell 1996). The cognitive interview is a method for gathering detailed information from respondents about the processes involved in the formulation of responses and involves extensive probing, either during the interview or immediately after it (e.g., Jobe, Tourangeau, and Smith 1993). Analysis of the verbal output based on "think aloud" protocols obtained from respondents retrospectively or while they answer questions has also been implemented, sometimes with automatic coding of the protocols (e.g., Bolton 1993). In addition, methods for coding the questionnaire itself have also been developed (e.g., Lessler and Forsyth 1996).

JOHN N. BASSILI is professor of psychology at the Scarborough campus of the University of Toronto. B. STACEY SCOTT is an undergraduate student at that campus. This research was supported by Social Sciences and Humanities Research Council of Canada grant 410-94-0170.

Public Opinion Quarterly Volume 60:390–399 0033-362X/96/6003-0003$02.50

The present research explores the utility of a cognitive index of information processing in the identification of question problems. The index is response latency, a measure that is increasingly simple to obtain in survey research with the widespread use of computers. For example, a previous work (Bassili 1996) discusses a methodology that allows accurate response latency measurement in CATI surveys that can be implemented more economically than the question-screening methods just reviewed.

Response latency is a general measure of the amount of information processing necessary to answer a question. The type of information processing indexed by response latency is diverse. To appreciate this, it is useful to examine a popular model of the steps involved in answering a question (Strack and Martin 1987; Tourangeau 1987; Tourangeau and Rasinski 1988). According to this model, question answering involves four distinguishable steps: question interpretation, memory retrieval, information integration, and response selection. Response latency is sensitive to information processing at each one of these steps (see Bassili 1996). Because question-answering problems can arise from difficulties at each of the four steps of the model, response latency provides a fittingly broad index of information-processing demands. The working assumption we adopt in this research is that question problems tend to slow responses because the resolution of the problem requires processing time. The specific aim of the research, therefore, is to test whether bad questions take longer to answer than good questions. To the extent that this is the case, response latency has the potential for signaling question problems.

The research explores response latency for three types of bad questions: questions containing a superfluous negative, questions containing a reference to two distinct themes (i.e., double-barreled questions), and questions that were shown by behavioral research to elicit high levels of problem behaviors (Fowler 1992). A "repaired" version of each question was prepared (the double-barreled questions had two repaired versions, one focusing on each theme) and, although the order of questions in the survey was fixed, the version of each question presented to a respondent was determined randomly. This approach allowed us to examine experimentally the impact of known question problems on response latency.

The Study

Two interviewers conducted the field work by telephone in the spring of 1994 from a laboratory facility at an urban university. A random sample of 289 valid student telephone numbers was drawn from the

university's registration records, from which 200 successful completions were secured, representing a response rate of 69.2 percent.

The questionnaire consisted of 26 opinion questions, three of which, in their bad form, contained a superfluous negative and six of which made reference to two distinct themes. These questions were answered by a simple expression of agreement or disagreement with the assertion contained in the item. Four questions that behavioral research (Fowler 1992) has shown to be problematic were also included. Of these, one was answered by "Yes" or "No," and the rest required frequency estimates. Each opinion question had a "good" form that did not contain a superfluous negative or that referred to only one of the two themes touched on by a double-barreled question. The good form of the four questions derived from behavioral research consisted of the repaired versions developed by Fowler (1992).

Response latency was measured by the interviewer by pressing the space bar of the computer keyboard upon finishing the delivery of the question and again as soon as the respondent gave an answer. Measured response latencies, therefore, do not include the time taken by the interviewer to read the question. Response latencies in which respondents asked for any type of information before answering were coded as invalid by the interviewer and were treated as missing data (for details on this procedure, see Bassili 1996). The coding of the validity of the response latency measure was followed by a secondary code tracking three behaviors: whether a request for repetition or clarification was made and whether, in the case of double-barreled questions, the respondent asked which aspect of the question should be answered. Because these problem behaviors involve questions to the interviewer, they always resulted in the invalidation of the corresponding response latencies.

The Findings

Latency invalidation occurred on an average of about 11 percent of the cases across questions. In addition, "Don't knows" were also excluded, and to minimize the impact of outliers in our analyses, latencies were truncated at two standard deviations above the mean.[1] Index

1. "Don't knows" averaged 3.8 percent and ranged from 0 percent to 11.5 percent, the latter figure coming from the double-barreled question on aboriginal rights. Because "Don't knows" were more frequent for one of the single-barreled forms of this question than for the double-barreled form, theme multiplicity does not appear to be responsible for this high figure. There were no refusals to answer among the present responses. Note also that to reduce the skewness of distributions of response latency scores, logarithmic or reciprocal transformations are often used. The reciprocal transformation of the present scores did not alter the substantive findings presented here.

Table 1. Mean Response Latencies and Proportion of Problem Behaviors for Questions with Superfluous Negatives and Their Affirmative Versions

Question Form	Mean Latency	Repeat or Clarify (%)
Negative: More should be done by businesses to reduce inequality in the workplace.	4.9[a]	6.3
Affirmative: More should be done by businesses to increase equality in the workplace.	4.8[a]	2.9
Negative: Policies that do not safeguard the environment are bad.	3.8[a]	15.7
Affirmative: Policies that safeguard the environment are good.	3.3[b]	4.1
Negative: Canada does not do enough to reduce pollution.	3.6[a]	4.1
Affirmative: Canada should do more to reduce pollution.	3.1[b]	1.0

NOTE.—Latencies are in seconds, rounded off to one decimal point. Means within a question type that do not share superscripts differ from each other with $p < .001$. Problem behaviors were coded by the interviewer after coding the validity of the response latency measure.

scores were prepared for each type of question by averaging response latencies, in standard scores, for the bad and the good forms of the questions, respectively.

SUPERFLUOUS NEGATIVES

A t-test for paired samples revealed that, on the average, questions took longer to answer when they contained a superfluous negative (M = 4.1 seconds) than when they did not (M = 3.8 seconds), $t(139) = 5.13$, $p < .001$. The differences were significant for two of the three questions (see table 1).

On the average, questions with a superfluous negative prompted requests for repetition or clarification 8.7 percent of the time, whereas affirmative questions did so only 2.7 percent of the time (see table 1). The slowing effect of superfluous negatives is thus consistent with expected behavioral consequences of problem questions.

DOUBLE BARRELS

A one-way repeated-measure ANOVA showed that the average response latencies for double-barreled items (M = 7.7 seconds) were longer than for questions containing only one theme (M = 5.2 seconds for the first theme and M = 6.7 for the second theme), F = 15.58, df = 2,244, $p < .001$. All six questions took significantly longer to answer when they contained two themes than when either of their themes was presented alone (see table 2). Theme multiplicity thus appears to slow responses.

On the average, respondents asked for clarification or repetition of the question, or for guidance on which "barrel" to answer, in 13.5 percent of the cases for double-barreled questions and in 11.4 percent of the cases for single-barreled questions (see table 2).

BEHAVIORAL QUESTIONS

The average response latency for the four questions that past behavioral research had shown to engender a high incidence of problem behaviors was compared to the average response latency for their repaired versions. These results contained a surprise. Although a t-test for paired samples was highly significant, $t(153) = 7.35$, $p < .001$, it is the repaired versions of these questions that took longer to answer than the problematic versions (M = 9.3 seconds vs. M = 7.7 seconds). The effect was observed in three of the four questions (see table 3).

As shown in table 3, the proportion of respondents who asked for the question to be repeated or clarified was higher, on the average, for the original version of questions (15.9) than for their repaired version (13.5). The relatively small difference in average percentages, however, hides the fact that in three of the four cases, Fowler's repaired versions were dramatically better than the original versions in minimizing requests for repetition and clarification. The one exception, which showed a marked reversal in this pattern, was the question about illness. Our presentation of this question, however, omitted transitional comments that were included by Fowler and that may have been critical to the clarity of the question.

Discussion

The results from questions containing a superfluous negative or a multiplicity of themes provide clear evidence for the slowing effect of poor questions on response latency. To appreciate this finding, it is important to bear in mind that timing in the present methodology starts

at the end of the question and does not include the time spent by the interviewer reading the question.[2] It appears, instead, that the extra time is spent by the respondent resolving matters having to do with the referent of the question. For questions that contain a superfluous negative, this extra time is most likely spent disentangling the meaning of the question. In the context of the four-step question-answering model discussed here, this activity is relevant to step 1 (question interpretation). In the case of double-barreled questions, the extra time is probably spent choosing a focus among the two options presented in the question and/or integrating feelings toward the two foci into one evaluation. These mental activities are thus relevant to the first (question interpretation) and third (information integration) steps of the question-answering model.

The results associated with the four questions derived from research involving behavior coding stand in contrast to those just discussed, not only because these questions took longer to answer in their repaired than in their problematic form but because response latencies were generally much longer than for the other questions in this study. We suspect that the pattern of results associated with these questions reflects the cognitive demands of frequency estimation (see Felcher and Calder 1990). Ironically, clear questions soliciting frequency estimates may focus respondents better on the demanding memorial search required by frequency estimation than questions that solicit the task with less clarity. Accordingly, differences in response latencies for these questions reflect activity at the second step (memory retrieval) of the question-answering model.

How can the relationships between question problems and response latency documented here be used in screening survey questions? The answer, we believe, must consider the broad context of information that is always necessary to the interpretation of response latencies. In the same way that scores derived from other methods for screening questions (e.g., behavior coding, protocol analysis, cognitive interviewing) require interpretation by comparison to known norms or variations in question form, so too do response latency scores. Thus, we suspect that researchers would benefit from the development of response latency norms for the types of questions they customarily include in their questionnaires.

It is instructive that many but not all the latency effects documented

2. Although timing was begun after the interviewer finished reading the question, the mere length of the question possibly still affected response latencies. Because affirmative and single-barreled questions are generally shorter than their problematic counterparts, and because the repaired versions of Fowler's questions are longer than the original versions, it would be advisable for future researchers to attempt to control for question length.

Table 2. Mean Response Latencies for Double-Barreled Questions and Their Two Single Barrels

Question Form	Mean Latency	Repeat or Clarify/Theme (%)
Double: It is important for Canada to make a special effort to protect ethnic and racial minorities and to mend the rift between Quebec and the rest of Canada.	18.3[a]	29.4
Single A: It is important for Canada to make a special effort to protect ethnic and racial minorities.	15.4[b]	11.9
Single B: It is important for Canada to mend the rift between Quebec and the rest of Canada.	15.2[b]	24.6
Double: Aboriginal peoples should not have to pay taxes, and their takeover of a downtown building was a justifiable action against the government.	8.3[a]	15.9
Single A: Aboriginal peoples should not have to pay taxes.	3.7[b]	6.8
Single B: The Aboriginal takeover of a downtown building was a justifiable action against the government.	7.7[c]	14.7
Double: Having an extended school year will help students focus on their studies more and keep them off the street and out of trouble.	7.4[a]	16.4
Single A: Having an extended school year will help students focus on their studies more.	5.0[b]	29.4

Single B: Having an extended school year will help keep students off the streets and out of trouble.	5.3[c]	11.7
Double: Same-sex couples and single parents should not be able to adopt children.	5.0[a]	8.6
Single A: Same-sex couples should not be able to adopt children.	3.9[b]	4.6
Single B: Single parents should not be able to adopt children.	3.6[c]	8.1
Double: The level of crime has increased over the last couple of years, and Toronto qualifies as a high-crime city.	6.5[a]	8.9
Single A: The level of crime has increased over the past couple of years.	3.9[b]	3.5
Single B: Toronto qualifies as a high-crime city.	3.2[c]	4.8
Double: If a patient suffers from a dangerous, contagious disease, the doctor should keep the patient's identity confidential and should only be required to notify local health authorities when there is a national health emergency.	12.7[a]	1.6
Single A: If a patient suffers from a dangerous, contagious disease, the doctor should only notify local health authorities when there is a national health emergency.	7.1[b]	9.5
Single B: If a patient suffers from a dangerous contagious disease, the doctor should keep the patient's identity confidential.	9.1[c]	7.3

NOTE.—Latencies are in seconds, rounded off to one decimal point. Means within a question type that do not share superscripts differ from each other with $p < .05$. Problem behaviors were coded by the interviewer after coding the validity of the response latency measure.

Table 3. Mean Response Latencies for Behavior-Coded Questions and Their Repaired Versions

Question Form	Mean Latency	Repeat or Clarify (%)
Original: Do you exercise or play sports regularly?	12.6[a]	5.0
Repaired: Do you do any sports or hobbies involving physical activities, or any exercise, including walking, on a regular basis?	17.1[b]	0.0
Original: What is the average number of days each week you have butter?	3.8[a]	30.1
Repaired: Not including margarine, what is the average number of days each week you have butter?	5.0[b]	9.7
Original: What is the number of servings of eggs in a typical day?	4.2[a]	11.7
Repaired: On days you eat eggs, how many eggs do you usually have?	3.9[b]	5.6
Original: During the past 12 months, that is, since January 1, 1994, about how many days did illness keep you in bed for more than half of the day?	8.8[a]	16.7
Repaired: During the past 12 months, since January 1, 1994, on about how many days did you spend several extra hours in bed because you were sick, injured, or just not feeling well?	10.1[b]	38.8

NOTE.—Latencies are in seconds, rounded off to one decimial point. Means within a question type that do not share superscripts differ from each other with $p < .001$.

here found their counterpart in behavior problems. For example, the last item in table 2 takes longer to answer in its double-barreled form than when either theme is presented alone, yet the double-barreled form produces fewest behavioral problems. This is probably because the two themes (reporting health problems to authorities vs. confidentiality) work together in creating a clear dilemma in the double-barreled form of the question, and this dilemma requires time to resolve. The

lesson for us is that long response latencies do not always indicate question problems.

Although the present research only begins to explore the relationship between response latency and other indexes of question problems, the results we have presented suggest that although overlap probably occurs in the indexing value of these measures, the overlap is not perfect. It would be interesting in the future to explore the distinctive contribution that response latency measurement can make to the identification of particular question problems (Presser and Blair 1994).

Response latency provides an unusually economical index of question problems, and we believe that researchers will benefit from tracking on an ongoing basis how long it takes respondents to answer their questions.

References

Bassili, John N. 1996. "The How and the Why of Response Latency Measurement in Telephone Surveys." In *Answering Questions*, ed. Norbert Schwarz and Seymour Sudman, pp. 319–46. San Francisco: Jossey-Bass.

Bolton, Ruth N. 1993. "Pretesting Questionnaires: Content Analyses of Respondents' Concurrent Verbal Protocols." *Marketing Science* 12:280–303.

Felcher, E. Marla, and B. Calder. 1990. "Cognitive Models for Behavioral Frequency Estimation." In *Advances in Consumer Research*, vol. 17, ed. M. Goldberg, G. Gorn, and R. W. Pollay. Provo, UT: Association for Consumer Research.

Fowler, Floyd. 1992. "How Unclear Terms Affect Survey Data." *Public Opinion Quarterly* 56:218–31.

Fowler, Floyd, and Charles Cannell. 1996. "Using Behavioral Coding to Identify Cognitive Problems with Survey Questions." In *Answering Questions*, ed. Norbert Schwarz and Seymour Sudman, pp. 15–36. San Francisco: Jossey-Bass.

Jobe, Jared, Roger Tourangeau, and Albert F. Smith. 1993. "Contributions of Survey Research to the Understanding of Memory." *Applied Cognitive Psychology* ("Special Issue: Practical Aspects of Memory: The 1994 Conference and Beyond") 7:567–84.

Lessler Judith T., and Barbara H. Forsyth. 1996. "A Coding System for Appraising Questionnaires." In *Answering Questions*, ed. Norbert Schwarz and Seymour Sudman, pp. 259–91. San Francisco: Jossey-Bass.

Presser, Stanley, and Johnny Blair. 1994. "Survey Pretesting: Do Different Methods Produce Different Results?" *Sociological Methodology* 24:73–104.

Strack, Fritz, and L. L. Martin. 1987. "Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys." In *Social Information Processing and Survey Methodology*, ed. H. J. Hippler, Norbert Schwarz, and Seymour Sudman, pp. 123–48. New York: Springer.

Tourangeau, Roger. 1987. *Evaluation of Data Collection Frequency and the Use of a Summary in the National Medical Care Utilization*. Hyattsville, MD: Public Health Service, National Center for Health Statistics.

Tourangeau, Roger, and Kenneth A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103:299–314.